Web Data - CS 483

Whoosh Indexing Report

Super Clever Team Name

Aaron Goin
Giselle Gomez
Richie Zhang

Mar 10th 2017

Introduction

For Assignment 3 our team was to create a search index using Whoosh! over our previously collected datasets and to allow searching the index from the command-line. Our dataset was obtained from Wikipedia and comprised of all stubs contained within the Confections, Desserts, and Fruit categories. We wrote one script to scrape and parse the xml data into a SQL database. We then wrote another script to index the data with Whoosh! without stemming the attributes. To make up for the lack of stemming, we use query.Variations in our Query Parser which calculates word variations for every term the user enters and broadens our search net.

Database

Schema

Our SQL database consisted of three tables with identical schemas:

Category
Name: STRING
Image: STRING
Description: STRING
Source: STRING

Our Whoosh! indexed database combined the three tables into a single index with a simple schema:

Index
Name: TEXT
Category: TEXT
Image: TEXT
Description: TEXT
Source: TEXT

Both Image and Source are string URLs

Size

Our database contains 1077 tuples in total, with 403 Desserts, 372 Sweets, and 302 Fruits.

Examples